Quality Control and Data Mining Techniques Applied to Monitoring Scaled Scores

A. A. VON DAVIER
Educational Testing Service, Princeton, NJ, USA

For testing programs that provide a large number of administrations each year, the challenge of maintaining comparability of test scores is influenced by the potential rapid accumulation of errors and by the lack of time between administrations to apply the usual techniques for detecting and addressing scale drift. Traditional quality control techniques have been developed for tests with only a small number of administrations per year, and therefore, while very valuable and necessary, they are not sufficient for catching changes in a complex and rapid flow of scores. Model-based techniques that can be updated at each administration could be used to flag any unusual patterns. The basis for the paper is recent research conducted at Educational Testing Service. I will describe an application of traditional quality control charts, such as Shewhart and CUSUM charts on testing data, time series models, change point models, and hidden Markov models to the means of scaled scores to detect abrupt changes. Some preliminary data mining approaches and results also will be discussed. This type of data analysis of scaled scores is relatively new and any application of the aforementioned tools is subject to the typical pitfalls: Are the appropriate variables included? Are the identified patterns meaningful? Can time series models or hidden Markov models be generalized to data from other tests?

Key Words and Phrases: data mining, quality control, scale drift, scaled scores, time series, Shewhart charts, CUSUM charts, change-point models, hidden Markov models

1. INTRODUCTION

In this paper I provide an overview of recent research conducted at Educational Testing Service (ETS) to enhance the data analysis, monitoring, classification, and prediction in evaluating equating results. The perspective I take here is that quality control and data mining tools from manufacturing, biology, and text analysis can be successfully applied to scaled scores and other relevant variables of an assessment. The quality control techniques may help with detecting trends, while the data mining tools may help with identifying (useful) patterns in the data that accompany the scaled scores.

In recent years at ETS, researchers considered monitoring the following variables: means and variances of the scaled and raw scores, means and variances of item parameters after they were placed on a common item response theory (IRT) scale, IRT linking parameters over time (the estimated slope and intercept of the linear relationship between the item/person parameters from the old and new administrations or from the item bank and the new administration), correlations among different sections of the tests, automatic and human scoring data, background variables, and so on. Some of these variables have been investigated by the team responsible for the quality of scores, but in the recent years, this investigation became more focused on patterns over a long chain of administrations. We attempted to address these inquires by using Shewhart control charts to visually inspect the data over time, time series models to model the relationship of test difficulty and test scores means over time, harmonic regression to remove seasonality, cumulative sum (CUSUM) charts, change-point models and hidden Markov models to detect sudden changes, weighted mixed models and analysis of variance to detect patterns in the data.

2. THE PROCESS OF QUALITY CONTROL IN ASSESSMENT

A brief exposition of the quality control process of the assessment data is as follows: After a testing administration and after the customary data analysis is conducted, the next step is to inspect Shewhart control charts for individual or average of the means of scaled scores over time. One visually inspects the control charts and identifies outliers. CUSUM charts should be inspected next.

The next step is applying time-series techniques to assess the level of autocorrelation and the degree of seasonality in the data as in Lee and Haberman (2011). Or one might model the series of individual raw-to-scale conversions over many administrations using a regression model with autoregressive moving-average (ARMA) errors (see Li, Li, & von Davier, 2011).

Then, one may consider applying a change-point model or a hidden Markov model to detect a point in time when the test results might contain a significant change (see Lee & von Davier, 2011). The main tasks of change-point detection are first to decide whether there has been a change, and if so, to estimate the time at which it occurred.

One might be interested in mining the data further by identifying patterns of test scores per subgroups of test takers. Luo, Lee, and von Davier (2011) investigated a multivariate weighted mixed model where the means of scaled scores are predicted by several background variables and the Test Administration variable, which is defined by specific sample compositions at each administration.

3. CONCLUSIONS

This paper presents a new perspective on quality control in assessments that is appropriate for the new generation of tests that have a continuous or almost continuous administration mode and that are delivered on the computer (and therefore, allow for the collection of additional information, such as time responses). These types of assessments include linear tests but also computer adaptive tests, multi-stage adaptive tests, and linear on-the-fly tests. Moreover, the tools described here can be applied to other assessment variables of interest. These tools can support the validity of the test overtime through timely identifying security breaches, administration errors, or demographic changes. As with all new applications, the approaches described here require more in-depth analyses to refine the approaches for matching the type of data from educational assessments. The theoretical and practical implications of the issues discussed in this paper are crucial for all standardized assessments with nontraditional equating designs and features.

REFERENCES

LEE, Y.-H., & HABERMAN, S. (2011). *Application of harmonic regression to monitor scale stability.* Manuscript in preparation.

LEE, Y.-H., & VON DAVIER, A. A. (2011, April). *Monitoring scale scores over time via quality control tools and time series techniques.* Paper presented at the annual meetings of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME), New Orleans, LA.

LI, D., LI, S., & VON DAVIER, A. A. (2011). Applying time-series analysis to detect scale drift. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 327–346). New York, NY: Springer-Verlag.

LUO, L., LEE Y.-H., & VON DAVIER, A. A. (2011). *Pattern detection for scaled score means of subgroups across multiple test administrations.* Paper presented at the annual meetings of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME), New Orleans, LA